UNIFORM: Unifying Knowledge from Large-scale and Diverse Pre-trained Models

Supplementary Material

In this technical Appendix, we present details on the experiments, including baselines and public teachers.

6. Experiments

6.1. How Well Does UNIFORM Scale - Number of Descriptive Teachers

We present the visualization of the performance of UNIFORM when scaling up the number of public descriptive teachers in Figure 6. The results are consistent with the findings in the main paper, demonstrating that UNIFORM can effectively leverage the knowledge from a large number of descriptive teachers to improve the student's performance. Though we found that the performance tends to saturate when the number of teachers exceeds 30 on some datasets, the performance continues to improve with more teachers on other datasets.

6.2. Details of Baselines

To compare the effectiveness of UNIFORM, we establish two types of baselines, detailed below:

Methods with predictive teachers only. We introduce three baselines, *i.e.*, Knowledge Distillation (KD) [22], CFL [37], and OFA [19] without ground-truth labels. These methods learn solely from predictive teachers without relying on ground-truth labels.

Knowledge Distillation (KD). In this baseline, we employ logit and feature transfer from homogeneous teachers, which are trained on the same dataset and share the same architecture as the student. Specifically, we denote the logits and features from N^{homo} teachers as $\{\boldsymbol{x}_{i,j}^t\}_{j\in[N^{homo}]}$ and $\{\boldsymbol{p}_{i,j}^t\}_{j\in[N^{homo}]}$ for the i-th input, respectively. The average feature $\bar{\boldsymbol{x}}_i^t$ and logit $\bar{\boldsymbol{p}}_i^t$ serve as the training targets for the i-th input,

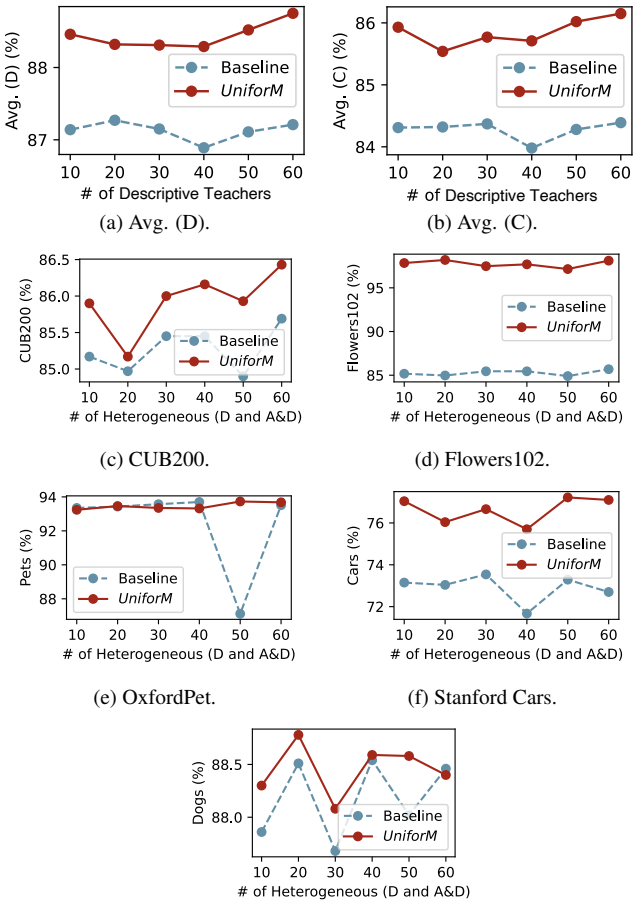
$$\begin{aligned}\bar{\boldsymbol{x}}_i^t &= \frac{\sum_{j\in[N^{homo}]}\boldsymbol{x}_{i,j}^t}{N^{homo}}\,,\\ \bar{\boldsymbol{p}}_i^t &= \frac{\sum_{j\in[N^{homo}]}\boldsymbol{p}_{i,j}^t}{N^{homo}}\,.\end{aligned} \tag{8}$$

The loss used in KD is

$$\ell_{KD} = dist(\boldsymbol{x}, \bar{\boldsymbol{x}}^t) + \sum_{c\in[C]} \bar{p}_c^t \log p_c\,, \tag{9}$$

where p_c is the c-th element of $\boldsymbol{p}$.

CFL. We adapt CFL to handle multiple teacher scenarios, incorporating all predictive teachers (both homogeneous and heterogeneous (A) types), learning from the average features and logits as KD. The hyperparameters remain consistent with those in the original paper.



(a) Avg. (D). (b) Avg. (C).

(c) CUB200. (d) Flowers102.

(e) OxfordPet. (f) Stanford Cars.

(g) Stanford Dogs.

Figure 6. Performance when scaling up the number of public descriptive teachers under the 5 datasets setting (main configuration).

OFA. For OFA, we modify its design by removing the cross-entropy loss from Equation (1) of the original paper while keeping all other details unchanged. OFA also learns from the average features and logits.

Methods with predictive and descriptive teachers: Since there does not exist a method that can learn from predictive and descriptive teachers simultaneously, we extend CFL to support the descriptive teachers as our baseline, referred to as CFL+.

CFL+ is specifically designed to accommodate descriptive teachers, *i.e.*, heterogeneous (D and A&D) teachers, and adopts the same encoder and decoder structure ($f^e(\cdot)$ and $f^d(\cdot)$) as UNIFORM. and $f^d(\cdot)$) as UNIFORM. However, different from UNIFORM, CFL+ directly averages the

features and logits from all public teachers to serve as the student's targets, similar to KD. Specifically, CFL+ employs the following loss,

$$\begin{aligned} \ell_{logit} &= \sum_{c\in[C]} \bar{p}_c^t \log p_c \,, \\ \ell_{feature} &= dist(\bar{\boldsymbol{x}}^t, \bar{\boldsymbol{x}}) \,, \\ \ell_{CFL+} &= \ell_{logit} + \beta_1 \ell_{feature} + \beta_2 \ell_{rec} \,, \end{aligned} \tag{10}$$

where β_1 and β_2 are the same with UNIFORM.

In all baseline methods, the label spaces of the logits are unified across different datasets. When combining multiple datasets, the label spaces of all datasets are merged into a single, unified space. This ensures that logits from all predictive teachers align with the same set of classes, enabling fair comparisons with UNIFORM, which also operates in a unified label space. Specifically, if there are K datasets, each with a label space $\mathcal{Y}_i = \{y_c\}_{c\in[C_i]}$, the **unified label space** $\mathcal{Y}$ is the union of all individual label spaces $\mathcal{Y} = \bigcup_{i=1}^{K} \mathcal{Y}_i$. Each sample is assigned a target in the combined label space, and logits from all predictive teachers are aligned with this unified set of classes. For a student learning from K datasets, the logits $\boldsymbol{p}_i^t$ from the i-th teacher are projected (reassigning indices and padding with zero) to this unified space, ensuring consistency across datasets.

6.3. Details of Public Teachers

The predictive and descriptive teachers used in our experiments are detailed in Tables 8 to 10.

For the **predictive teachers**, we select four base models per dataset, resulting in a total of 44 predictive teachers across the 11 datasets, as listed in Table 8.

For the **descriptive teachers**, we aim to leverage both well-known representative models and widely used public models. Specifically:

- We include 10 representative models chosen based on expert knowledge, as detailed in Table 9.
- Additionally, we select the top 50 most downloaded models from HuggingFace, as shown in Table 10.

In total, for the experiments conducted on 11 datasets, we utilize $44 + 10 + 50 = 104$ teachers.

Dataset (Model)	HuggingFace Model
CUB200 (ResNet-50)	anonauthors/cub200-resnet50
CUB200 (ViT-b32)	anonauthors/cub200-ViT-b32
CUB200 (ConvNeXt-base)	anonauthors/cub200-ConvNeXt-base
CUB200 (SwinTransformer)	anonauthors/cub200-swinT
Flower102 (ResNet-50)	anonauthors/flowers102-resnet50
Flower102 (ViT-b32)	anonauthors/flowers102-ViT-b32
Flower102 (ConvNeXt-base)	anonauthors/flowers102-ConvNeXt-base
Flower102 (SwinTransformer)	anonauthors/flowers102-swinT
Oxford Pets (ResNet-50)	anonauthors/oxford_pet-resnet50
Oxford Pets (ViT-b32)	anonauthors/oxford_pet-ViT-b32
Oxford Pets (ConvNeXt-base)	anonauthors/oxford_pet-ConvNeXt-base
Oxford Pets (SwinTransformer)	anonauthors/oxford_pet-swinT
Stanford Cars (ResNet-50)	anonauthors/stanford_cars-resnet50
Stanford Cars (ViT-b32)	anonauthors/stanford_cars-ViT-b32
Stanford Cars (ConvNeXt-base)	anonauthors/stanford_cars-ConvNeXt-base
Stanford Cars (SwinTransformer)	anonauthors/stanford_cars-swinT
Stanford Dogs (ResNet-50)	anonauthors/stanford_dogs-resnet50
Stanford Dogs (ViT-b32)	anonauthors/stanford_dogs-ViT-b32
Stanford Dogs (ConvNeXt-base)	anonauthors/stanford_dogs-ConvNeXt-base
Stanford Dogs (SwinTransformer)	anonauthors/stanford_dogs-swinT
Food101 (ResNet-50)	anonauthors/food101-resnet50
Food101 (ViT-b32)	anonauthors/food101-ViT-b32
Food101 (ConvNeXt-base)	anonauthors/food101-ConvNeXt-base
Food101 (SwinTransformer)	anonauthors/food101-swinT
Caltech101 (ResNet-50)	anonauthors/caltech101-resnet50
Caltech101 (ViT-b32)	anonauthors/caltech101-ViT-b32
Caltech101 (ConvNeXt-base)	anonauthors/caltech101-ConvNeXt-base
Caltech101 (SwinTransformer)	anonauthors/caltech101-swinT
Cifar10 (ResNet-50)	anonauthors/cifar10-resnet50
Cifar10 (ViT-b32)	anonauthors/cifar10-ViT-b32
Cifar10 (ConvNeXt-base)	anonauthors/cifar10-ConvNeXt-base
Cifar10 (SwinTransformer)	anonauthors/cifar10-swinT
Cifar100 (ResNet-50)	anonauthors/cifar100-resnet50
Cifar100 (ViT-b32)	anonauthors/cifar100-ViT-b32
Cifar100 (ConvNeXt-base)	anonauthors/cifar100-ConvNeXt-base
Cifar100 (SwinTransformer)	anonauthors/cifar100-swinT
DTD (ResNet-50)	anonauthors/dtd-resnet50
DTD (ViT-b32)	anonauthors/dtd-ViT-b32
DTD (ConvNeXt-base)	anonauthors/dtd-ConvNeXt-base
DTD (SwinTransformer)	anonauthors/dtd-swinT
Aircraft (ResNet-50)	anonauthors/fgvc_aircraft-resnet50
Aircraft (ViT-b32)	anonauthors/fgvc_aircraft-ViT-b32
Aircraft (ConvNeXt-base)	anonauthors/fgvc_aircraft-ConvNeXt-base
Aircraft (SwinTransformer)	anonauthors/fgvc_aircraft-swinT

Table 8. Online available public predictive teachers.

#	Model Name
1	VGG19
2	ResNet18
3	ResNet34
4	ResNet50
5	ResNet101
6	ConvNext-Base
7	ViT-B32
8	DeIT-S16
9	ResMLP-12
10	SwinTransfromer-Base

Table 9. Selected representative public descriptive teachers.

Type	#	Model Name and HuggingFace Model
	1	timm/resnet50.a1_in1k
	2	google/vit-base-patch16-224
	3	timm/mobilenetv3_large_100.ra_in1k
	4	microsoft/beit-base-patch16-224-pt22k-ft22k
	5	timm/resnet18.a1_in1k
	6	amunchet/rorshark-vit-base
	7	rizvandwiki/gender-classification
	8	timm/convnext_small.fb_in22k
	9	nvidia/mit-b0
	10	timm/resnet18.fb_swsl_ig1b_ft_in1k
	11	trpakov/vit-face-expression
	12	timm/nfnet_l0.ra2_in1k
	13	timm/vit_tiny_patch16_224.augreg_in21k_ft_in1k
	14	timm/swin_base_patch4_window7_224.ms_in22k_ft_in1k
	15	nateraw/vit-age-classifier
	16	facebook/convnext-tiny-224
	17	NTQAI/pedestrian_gender_recognition
	18	Kaludi/food-category-classification-v2.0
	19	nateraw/food
	20	timm/vgg19.tv_in1k
	21	timm/vit_base_patch16_224.augreg_in21k
	22	timm/tf_efficientnet_b0.ns_jft_in1k
	23	timm/mobilenetv2_100.ra_in1k
	24	microsoft/dit-base-finetuned-rvlcdip
	25	apple/mobilevit-small
	26	timm/resnetv2_50x1_bit.goog_in21k
	27	aaraki/vit-base-patch16-224-in21k-finetuned-cifar10
	28	Falconsai/nsfw_image_detection
	29	timm/resnet101.a1h_in1k
	30	timm/efficientnet_b0.ra_in1k
	31	google/mobilenet_v2_1.0_224
	32	WinKawaks/vit-tiny-patch16-224
	33	timm/fbnetc_100.rmsp_in1k
	34	timm/tf_efficientnetv2_s.in21k
	35	google/mobilenet_v1_0.75_192
	36	timm/convnext_base.fb_in22k_ft_in1k
	37	microsoft/beit-base-patch16-224
	38	timm/vit_small_patch16_224.augreg_in21k_ft_in1k
	39	timm/hrnet_w18.ms_aug_in1k
	40	timm/mobilevit_s.cvnets_in1k
	41	timm/convnextv2_atto.fcmae_ft_in1k
	42	timm/cspdarknet53.ra_in1k
	43	microsoft/swin-tiny-patch4-window7-224
	44	timm/ese_vovnet19b_dw.ra_in1k
	45	timm/tf_efficientnetv2_s.in21k_ft_in1k
	46	timm/mixer_b16_224.goog_in21k_ft_in1k
	47	timm/deit_base_distilled_patch16_224.fb_in1k
	48	timm/pnasnet5large.tf_in1k
	49	timm/pit_b_224.in1k
	50	timm/mnasnet_100.rmsp_in1k

Table 10. Top 50 downloaded vision models on HuggingFace we use for public models. This might be different from HuggingFace as time changes.